Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Görüntü Kazıma Yoluyla Oluşturulan Örnek Veri Kümesinin Evrişimsel Sinir Ağı Tabanlı Görüntü Sınıflama Üzerine Etkisinin İncelenmesi

Tolga HAYIT [a, *]

[a]*Bilgisayar Mühendisliği Bölümü, Mühendislik-Mimarlık Fakültesi, Yozgat Bozok Üniversitesi, Yozgat, TÜRKİYE*

** Sorumlu yazarın e-posta adresi: tolga.hayit@bozok.edu.tr*

DOI:10.29130/dubited.1120967

ÖZ

Derin öğrenme tabanlı görüntü sınıflandırma çalışmalarının en önemli aşamalarından biri veri elde etme aşamasıdır. Modeli eğitecek veri setinin göreve özgü ve uygun kalitede olması gerekmektedir. Bu nedenle veri setinin oluşturulma süreci araştırmacılar için zahmetli ve yorucu bir süreç olabilmektedir. Web kazıma teknikleri çalışmalarda kullanılabilecek uygun veri setlerinin oluşturulmasında araştırmacılara çözümler sunmaktadır. Özellikle derin öğrenme gibi çok sayıda veri ihtiyacı bulunan görevlerde bu tekniklerin kullanılması süreci ciddi anlamda hızlandırabilmektedir. Bu bağlamda bu çalışma, örnek bir görüntü sınıflandırma görevi için görsel kazıma teknolojisi ile oluşturulan veri setinin sınıflandırmaya başarısını araştırmaktadır. Çalışmada farklı CNN modelleri kullanılarak, oluşturulan örnek veri seti eğitilmiştir. Test veri seti doğruluk değeri (%98,8) ve diğer performans ölçütleri görsel kazıma yoluyla elde edilen veri setinin görüntü sınıflandırma görevleri için kullanılabileceğini desteklemektedir.

Anahtar Kelimeler: *Görsel kazıma, Web kazıma, Evrişimsel Sinir Ağı, Derin Öğrenme, Görüntü sınıflandırma*

An investigation of the Effect of Dataset Sample Created via Image Scraping on Convolutional Neural Network Based Image Classification

ABSTRACT

Data acquisition is one of the most important stage of deep learning based image classification studies. The training dataset of the model should be task-specific and qualified. Therefore, the dataset creating can be exhausting and laborious process for researchers. Web scraping method offer solutions to researchers in creating suitable data sets that can be used in studies. The use of these techniques can accelerate the process, especially in tasks that require more data, such as deep learning. In this context, this study investigates the classification success of the data set created via image scraping for image classification tasks. In the study, the dataset was trained by using different Convolutional Neural Network models. Test accuracy (98,8%) and other performance metrics support that the dataset created via image scraping can be used for image classification tasks.

Keywords: *Image scraping, Web scraping, Convolutional Neural Network, Deep learning, Image classification*

Geliş: 25/05/2022, Düzeltme: 27/07/2022, Kabul: 30/09/2022

I. GİRİŞ

Veri, makine öğrenimi, örüntü tanıma, veri madenciliği gibi yapay zekaya dayalı birçok disiplin için araştırmanın önemli bir parçasıdır. Makine öğrenimi alanında daha fazla veriyi bulundurmak, bir makine öğrenimi modelinin bunu anlaması ve karılaştığı yeni bir veri için doğru tahminler yapabilmesi için yüksek bir şans sağlamaktadır. Bununla birlikte veri sayısına paralel olarak güvenilir ve kaliteli verilerle çalışılması modeli daha sağlam hale getirmektedir. Veri oluşturmak ya da verinin elde edilmesi aşaması, modele sunulacak veri setinin ilk hazırlık aşamasını oluşturmaktadır. Doğru veri setinin elde edilmesi, verinin hazırlanması, verilerin temizliği vb. gibi problemlerden daha öncelikli bir konumdadır.

Web teknolojisi, birçok disiplinde araştırmacılar için önemli bir bilgi kaynağıdır. Web, farklı formatlarda çeşitli kaynaklardan yararlı ya da yararsız yapılandırılmış ve yapılandırılmamış bilgiler içermektedir [1]. Kaggle, UCI Machine Learning Repository gibi paylaşıma açık sistemler veri toplamak için araştırmacılara alternatif çözümler üretse de araştırmacılar birden fazla web sitesinden veri toplamak ve verileri analiz etmek isteyebilir. Bu durumda araştırmacıların karşısına farklı problemler çıkmaktadır. Elde edilecek veriler tek bir web sitesinde de bulunsa örneğin aynı sayfada yer almayabilir ya da veriler araştırmacının beklediği kalitede ve formda olmayabilir. Ek olarak, Birçok web sitesi verileri bir buton ya da araç vasıtası ile araştırmacının kendi bilgisayarına depolamasına ya da indirmesine izin vermez [2]. Araştırmacı bu durumda verileri manuel olarak kopyalayıp kendi bilgisayarına indirmektedir. Bu işlem, özellikle derin öğrenme gibi algoritmalar için fazla veri gerekeceğinden dolayı çok maliyetli, yorucu ve zaman alıcı bir iştir.

Otomatik web kazıma, web sitesi ortamlarından yapılandırılmamış ya da işlenmemiş verileri otomatik olarak elde etmek için (çıkarmak) kullanılan bir tekniktir. Çıkarılacak verinin türü yapılacak çalışmaya göre görsel, işitsel, metin ya da sayısal olabilir. Web kazımada genel olarak bir web sitesinden veri çıkarmak ve bu veriyi yapılandırarak kullanılabilir ya da anlaşılabilir bir formata dönüştürmektir (Şekil 1).

Şekil 1. *Web kazıma genel yapısı*

Web kazıma teknikleri, klasik olarak kopyala-yapıştır gibi insan çabası gerektiren maliyetli ve zaman alıcı geleneksel çözümlerden verileri yapılandırarak indirebilen tam otomatik sistemlere kadar geniş bir yelpazeye sahiptir [3]. Bunların içerisinde makine öğrenimi ve özellikle bilgisayarlı görü alanlarında web sayfalarını bir insanın yapabileceği gibi görsel açıdan yorumlayarak sayfalardan verileri çıkarmaya çalışan uygulamalar da bulunmaktadır [4]. Bu uygulamalar "Görsel Kazıyıcı" olarak adlandırılmaktadır.

Görsel Kazıyıcı uygulamalarının kullanılması, etik açıdan bazı tartışmaları da beraberinde getirmiştir. Elde edilen verilerin gelişi güzel sergilenmesi telif açısından sıkıntı doğurabilmektedir. Google, Yandex, Bing vb. arama motorlarının görsel arama tarafı, görsel web kazıma uygulama örneklerinin başında gelmektedir. Bu uygulamalardan özellikle Google Görseller için sonuçlanan geçmişe dönük emsal davalar Google Görsellerin resimleri sergilemesinin "Adil Kullanım" bağlamında sorun teşkil etmediği yönünde sonuçlanmıştır [5-8]. Bununla birlikte Google Görseller uygulaması, ilgili web kaynaklarındaki resimleri birebir olmadan; boyutları ve kalitesini düşürerek yayınlamakta ve ilgili resim üzerinden resmin bağlı bulunduğu web kaynağına yönlendirmektedir. Resimleri belirli bir ücret karşılığı satan ortamların ön izlemeleri "filigran" kullanılarak Google Görsellerde sergilenmektedir. Resimler üzerinden ilgili siteye yönlendirme yapılması da bu firmaların lehine bir durumdur.

Hangi web kazıma tekniği kullanılırsa kullanılsın, yaklaşımların ana hedefi web sitelerinden verileri otomatik olarak yakalayarak daha yapılandırılmış bir halde kullanılmasını sağlamaktır [9]. Literatürde istatistik [10,11], tıp [12-16], gıda [17,18] ve tekstil [19] gibi çeşitli alanların yanı sıra veri madenciliği ve makine öğrenimi alanlarında son dönemde yapılan çalışmalar [20-24], veri kazıma tekniklerinin başarısını ortaya koymaktadır.

Diğer yandan, son zamanlarda, insan beynini temel alan derin öğrenme yaklaşımları kullanılarak görüntü sınıflandırma ve görüntü tanıma görevlerinde inanılmaz ilerlemeler kaydedilmiştir [25-31]. Bir derin öğrenme modeli olan Derin Evrişimsel Sinir Ağları (CNNs: Convolutional Neural Networks) [32], ham verilerle eğitilerek özellikle görüntü sınıflandırma ve nesne tespit görevleri için son teknoloji bir model haline ulaşmıştır. CNN'ler etiketlendirilmiş renkli ham verileri ön işlemeye gerek kalmadan başarılı bir şekilde sınıflandırabilmektedir [33]. Bununla birlikte CNN'ler bazı kısıtlamaları beraberinde getirmektedir. Görsel verilerin eğitimi için sayıca oldukça fazla ve kaliteli veriye ihtiyaç duyulmaktadır.

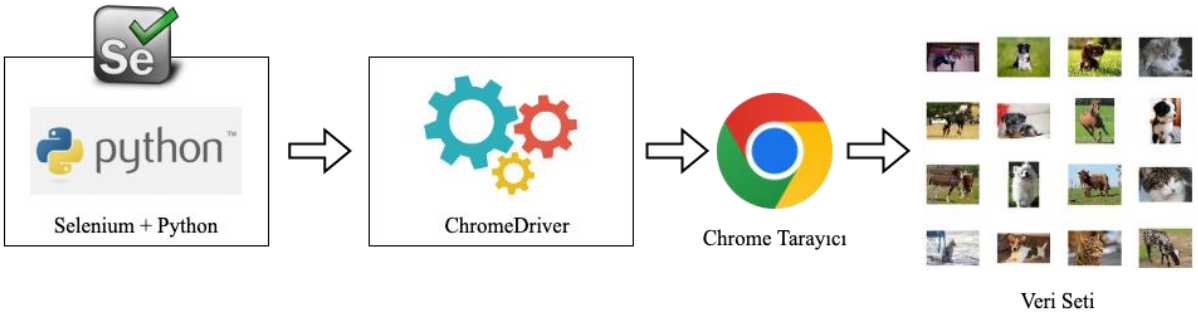
Bu çalışmada otomatik görsel web kazıma tekniği kullanılarak elde edilen veri setinin CNN tabanlı görüntü sınıflandırma modellerine etkisi incelenmiştir. Veriler dört farklı hayvan kategorisinde Google Görseller üzerinden otomatik olarak indirilerek oluşturulmuştur. Veri setinin eğitimi için görüntü sınıflandırma kategorisinde başarısı kanıtlanmış üç farklı CNN ağı işe koşulmuştur. Makalenin geri kalanı şu şekilde organize edilmiştir: sonraki bölümde veri seti, CNN ağları ve metodoloji, üçüncü bölümde elde edilen bulgularla birlikte karşılaştırmalı performans analiz sonuçları ve son bölümde de çalışmanın genel sonuçları sunulmuştur.

II. MATERYAL VE METOT

A. VERİ SETİ

Bir bilgisayarlı görü projesinin temel adımlarından biri veri elde etme ve etiketleme adımlarıdır. Veri seti oluşturma bu aşamanın en zahmetli ve zor kısmını oluşturmaktadır. Özellikle derin öğrenmeye dayalı bir görüntü sınıflama sistemi tasarlanıyorsa çok fazla etiketlendirilmiş veriye ihtiyaç bulunmaktadır. Çalışmanın bu aşamasında web kazıma tekniği kullanılarak, Google Görseller görsel arama motoru üzerinden veri seti oluşturulmuştur. Google Görseller, çevrimiçi görsel aramak için Google'ın sunduğu herkese açık web tabanlı son teknoloji bir ara yüzdür. Google Görsellere alternatif farklı ara yüzler de bulunmaktadır (Bing görseller, yandex görseller vb.). Hem bu alanda öncü olması bakımından hem de dünyada en çok kullanılan arama motoru olması bakımından veri seti oluşturmada Google Görseller kullanılmıştır. Google görseller üzerinden normal bir insanın manuel olarak resimleri indirip kaydetmesi oldukça zahmet gerektiren bir iştir. Bu nedenle bu işi bir insan yerine otomatik olarak yapabilecek bir ön sistem geliştirilmiştir.

Google görseller üzerinden görselleri otomatik ve hızlı bir şekilde indirebilmek için Python kodlama ile Selenium paketi kullanılmıştır. Selenium, web tarayıcılarının otomasyonunu sağlayan ve destekleyen bir dizi araç ve eklenti içeren; tarayıcı yazılımlarda kullanıcı etkileşimlerini taklit etmek için kod yazmaya izin veren ve altyapı sağlayan bir çatıdır [34]. Selenium, tarayıcı sürücüleri (web driver) üzerinden ara yüzlerine erişerek gerçek bir kullanıcının yeteneklerini simule etmektedir. Chrome sürücüsü (ChromeDriver) Selenium gibi uygulamaların tarayıcılar üzerinde işlem yapmasını sağlayan açık kaynaklı bir araçtır [35]. Web sayfalarında gezinme, form girişleri, JavaScript çalıştırma gibi birçok aktivite için olanaklar sunmaktadır. Chrome sürücüsü W3C standartlarını uygulayan bağımsız bir sürücüdür. Mac, Linux, Windows ve ChromeOS gibi farklı birçok işletim sistemi için kullanılabilmektedir. Python+Selenium ile Chrome sürücüsü kullanılarak toplamda dört kategoride (kedi, köpek, at ve inek) görsel verileri otomatik olarak indirilmiştir. Eğitim aşamasında sorun olmaması açısından indirilen veriler nihai olarak gerçek bir kullanıcı tarafından kontrol edilmiştir. Çalışmanın veri seti oluşturma iş akışı Şekil 2 üzerinde gösterilmiştir.

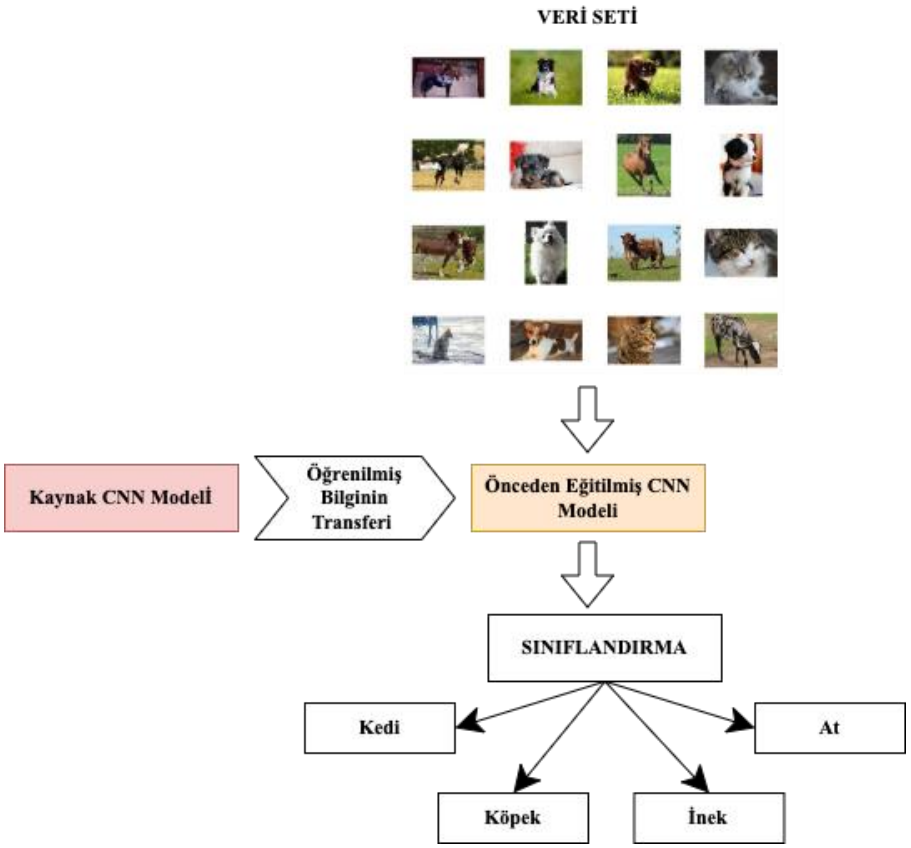


Şekil 2. *Veri seti iş akışı*

B. EVRİŞİMSEL SİNİR AĞLARI VE TRANSFER ÖĞRENİMİ

Evrişimsel sinir ağları (CNN'ler) görüntü, ses, metin gibi verilerden belirleyici özellikleri otomatik olarak öğrenebilen, çalışma biçimini insan beyninden ilham alan, birbirine bağlı katmanlardan oluşan bir yığın yapısıdır. Geleneksel yapay sinir ağlarına nazaran farklı görevlere ilişkin karmaşık ve doğrusal olmayan ilişkileri deneyimlemede ve tahmin etmede insanlarla aynı performansı gösterebilmektedir. Bir CNN mimarisi görüntü sınıflandırma görevi için sıfırdan eğitilebilmektedir; ancak bu süreç zaman alıcı ve zorlu bir süreçtir. Bu problemin önüne geçebilmek için mevcut görev için kullanılacak veri setinden daha büyük ölçekli veri seti kullanılarak önceden eğitilmiş modeller, veri seti ile birlikte sisteme girdi olarak sunulmaktadır. Transfer öğrenimi olarak adlandırılan ve giderek daha popüler hale gelen bu teknik kullanılarak daha az eğitim süreci geçirerek daha yüksek performans elde edilebilmektedir [36]. Literatürde MobileNet [37], ResNet [38], Inception (GoogleNet) [39] ve DenseNet [40] gibi önceden eğitilmiş birçok CNN modeli bulunmaktadır ve bunların her biri görüntü sınıflandırma görevi için kullanılmaktadır.

Bu çalışmada, görsel kazıma yöntemi ile elde edilen verisetinin farklı CNN modelleri kullanılarak sınıflandırma başarısına etkisi araştırılmıştır. Farklı sınıfların tanınabilmesi için önceden eğitilmiş InceptionV3, ResNet-101 ve DenseNet-201 modelleri işe koşulmuştur. Modeller, ImageNet projesi tarafından hemen her yıl düzenlenen ve farklı CNN sınıflandırıcıların performanslarının sergilendiği ImageNet Büyük Ölçekli Görsel Tanıma Yarışması (ILSVRC: ImageNet Large Scale Visual Recognition Competition) kapsamında başarı kaydeden bazı modellerin son versiyonları olacak şekilde seçilmiştir. Modellerin eğitilmesi, doğrulanması ve test edilmesi MATLAB programı kullanılarak gerçekleştirilmiştir. İşlemler sırasında Intel(R) Xeon(R) E-2236 marka CPU, 64 GB RAM ve ekran kartı olarak NVIDIA Quadro RTX A4000 donanımına sahip iş istasyonu kullanılmıştır. Önceden eğitilmiş her modelin yapısı dört kategori için yeniden tasarlanarak her model 4200 görüntü kullanılarak eğitilmiştir. Modeller eğitim sırasında 1200 görüntü ile doğrulanmıştır. Görsel kazıma ile oluşturulan veri seti kullanılarak oluşturulan CNN tabanlı sistemin iş akışının ana adımları Şekil 3'te gösterilmektedir.



Şekil 3. *CNN iş akışı*

C. PERFORMANS ÖLÇÜTLERİ

Bu çalışmaya benzer şekilde temelde çok sınıflı bir sınıflandırma görevi içeren çalışmalarda, CNN tabanlı bir modelin doğrulama ve test başarısı literatürde belirtilen bazı ölçütlere göre yapılmaktadır. Bu ölçütlerin başında doğruluk ölçütü gelmektedir. Ancak son çalışmalarda sınıflandırma başarısının performansının belirlenmesinde doğruluk ölçütünün tek başına yeterli olmadığı; başarının farklı ölçütlerle desteklenmesi gerektiği vurgulanmaktadır. Bir sınıflandırma görevi için karışıklık (hata) matrisi basitçe, denetimli bir öğrenme modelinin performansını görselleştiren bir tablodur. Tablonun her satırı sınıfların temsilini oluştururken; her sütun sınıflara karşılık gelen tahmin edilen örneklerin temsilini oluşturmaktadır. Çalışmanın başarısının daha sağlıklı bir şekilde belirlenebilmesi için karışıklık matrisi kullanılarak F1 Skor, hassasiyet ve kesinlik gibi farklı ölçütler hesaplanarak dikkate alınmıştır.

III. BULGULAR VE TARTIŞMA

Bu çalışmada Google Görseller üzerinden otomatik görsel kazıma yoluyla veri seti oluşturulmuştur. Python+Selenium ile Chrome Sürücüsü kullanılarak toplamda dört kategoride (kedi, inek, köpek ve at) görsel verileri otomatik olarak indirilmiştir. Oluşturulan veri setine ilişkin örnek görseller Şekil 4 üzerinde sunulmuştur.

Şekil 4. *Veri setinden örnek görseller*

Veriler hiçbir ön işleme işlemine tabi tutulmaksızın olduğu gibi ham haliyle kullanılmıştır. Veri seti istatistiklerine göre toplam veri sayısı modellerin eğitimi için yeterli seviyededir (Tablo 1). Bununla birlikte eğitimde aşırı uyum ya da yetersiz uyum gibi problemlerle karşılaşmamak adına; indirme işlemleri tamamlandıktan sonra yapılan kontroller sonucu ilgili sınıfı doğru temsil etmeyen görseller silinerek her bir sınıf için veri sayısı sabit tutulmuş ve bu sayede sınıflar arası tutarsız veri sorunu engellenmiştir. Oluşturulan nihai veri seti eğitim (%70), doğrulama (%20) ve test (%10) olmak üzere üç gruba ayrılmıştır. Nihai veri seti istatistikleri Tablo 1 üzerinde gösterilmiştir.

Tablo 1. *Her sınıf için toplam veri sayıları*

	Cat (Kedi)	**Cow** (İnek)	**Dog** (Köpek)	**Horse** (At)	**TOPLAM**
Eğitim (%70)	1050	1050	1050	1050	4200
Doğrulama (%20)	300	300	300	300	1200
Test (%10)	150	150	150	150	600
				TOPLAM	**6000**

Veri seti önceden eğitilmiş farklı CNN modelleri ile eğitilmiştir. Kullanılan CNN modelleri her bir model için en son ve en başarılı sürüm olacak şekilde belirlenmiştir. Eğitim süresince her bir model için elde edilen ortalama doğruluk değerleri Tablo 2'de gösterilmiştir. Buna göre CNN modelleri her sınıfı ortalama %97'nin üzerinde bir başarıyla sınıflandırabilmiştir. Adil bir süreç için tüm modellerin hiper parametre değerleri sabit tutulmuştur. Belirlenen bazı hiper parametre değerleri Tablo 3 üzerinde verilmiştir. Optimize edici olarak "Adam" algoritması kullanılmış ve maksimum eğitim turu (epoch) 20 olarak belirlenmiştir. Eğitimde her veri seti her turda karıştırılarak 50 doğrulama frekansı uygulanmıştır.

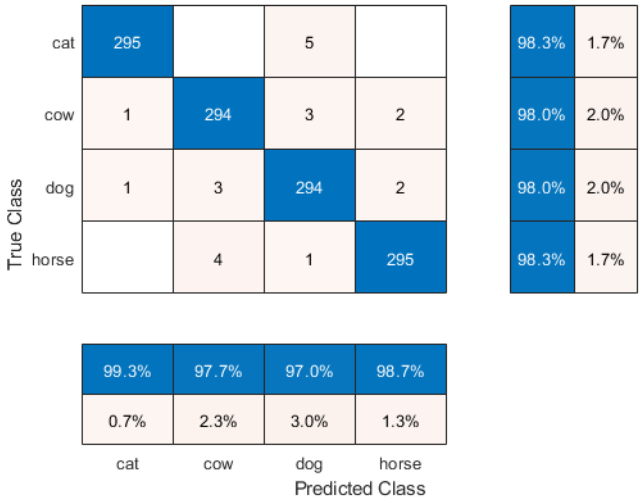
Tablo 2. *Her bir CNN modeli için ortalama doğruluk değerleri*

Model	**Doğruluk**
InceptionV3	%97
ResNet-101	%97
DenseNet-201	**%98**

Tablo 3. *CNN modelleri için kullanılan bazı hiper parametre değerleri*

Optimize Edici	Adam
Öğrenme Hızı	1e-05
Doğrulama Frekansı	50
Maksimum Eğitim Turu	20
Minimum Parti Boyutu	8
Karıştırma	Her turda

Diğer yandan, sınıflandırıcının her sınıfta nasıl performans gösterdiğini anlayabilmek için en başarılı model olan DenseNet-201 sonuçları üzerinden karışıklık matrisi oluşturulmuştur (Şekil 5). Karışıklık matrisi her sınıf için sınıflandırma başarısını ortaya koymaktadır. Satırlar gerçek sınıfları, sütunlar ise tahmin edilen sınıfları göstermektedir. Çapraz hücreler ise her sınıfın ne kadar doğru sınıflandırıldığını belirtmektedir. Karışıklık matrisi aracılığıyla sınıf bazında doğruluk, kesinlik, hassasiyet ve F1 Skor değerleri hesaplanmıştır (Tablo 4). Karışıklık matrisi ve bahsi geçen ölçütler, görsel kazıma yoluyla elde edilen veri seti kullanılarak oluşturulan CNN tabanlı görüntü sınıflandırma modelinin her sınıf için kabul edilebilir bir başarı ortaya koyduğunu göstermektedir.



Şekil 5. *DenseNet-201 modeli için karışıklık matrisi. Satırlar yukarıdan aşağıya; sütunlar soldan sağa doğru: cat (kedi), cow (inek), dog (köpek) ve horse (at)*

Tablo 4. *DenseNet-201 modelinin sınıf bazında tüm ölçütler için başarısı*

	Cat (Kedi)	**Cow** (İnek)	**Dog** (Köpek)	**Horse** (At)
Doğruluk (%)	99,33	98,42	98,5	98,75
Kesinlik	0,99	0,97	0,97	0,97
Hassasiyet	0,99	0,97	0,97	0,98
F1 Skor	0,99	0,97	0,97	0,97

Ek olarak daha sağlıklı bir sonuç elde etmek ve modellerin ezbere kaçmadığından emin olmak için; her bir model eğitim süresince karşılaşmadığı 600 adet görüntü ile ayrıca test edilmiştir. Her bir model için elde edilen test-doğruluk oranı Tablo 5 üzerinde gösterilmiştir. Test veri kümesinden elde edilen doğruluk oranlarının eğitim-doğrulama doğruluk oranlarına yakın hesaplanması, modellerin ezberleme, aşırı-yetersiz uyum gibi problemlerden etkilenmediğini ortaya koymaktadır. Eğitim ve test sonuçları modellerin kullanılan veri setine iyi bir şekilde genelleştirildiğini göstermektedir.

Tablo 5. *Her model için ortalama test-doğruluk değerleri*

Model	**Doğruluk**
InceptionV3	**%98,8**
ResNet-101	%96,8
DenseNet-201	**%98,8**

IV. SONUÇ VE ÖNERİLER

Bu çalışma Evrişimsel Sinir Ağı (CNN) tabanlı bir görüntü sınıflandırma modeli için görsel kazıma tekniği ile oluşturulan veri setinin kullanılabileceğini göstermiştir. Oluşturulan veri setinin kullanıldığı farklı CNN ağlarının performansları karşılaştırılmıştır. Doğruluk (%97,3) ve diğer tüm performans ölçütleri oluşturulan veri setinin her bir sınıfının başarılı bir şekilde sınıflandırıldığını göstermiştir. En yüksek doğruluk değeri sağlayan DenseNet-201 modeli test veri kümesini %98,8 bir doğrulukla sınıflandırabilmiştir. Bu sonuçlar doğrultusunda sonraki çalışmalarda görsel kazıma tekniği ile oluşturulan farklı veri kümelerinin görüntü sınıflama problemlerinde kullanılması önerilmektedir. Bununla birlikte modern bir yaklaşım olarak derin öğrenme yaklaşımlarının geleneksel görüntü sınıflama yaklaşımlarına göre performansının daha iyi seviyede olduğu literatürde belirtilse de görsel kazıma ile oluşturulan veri setleri üzerinde modern ve geleneksel yaklaşımların performanslarının karşılaştırılması ve sonuçların değerlendirilmesi önerilmektedir.

V. KAYNAKLAR

[1] R. Diouf, E. N. Sarr, O. Sall, B. Birregah, M. Bousso, and S. N. Mbaye, "Web scraping: state-of-the-art and areas of application," *IEEE International Conference on Big Data (Big Data)*, 2019, pp. 6040-6042.

[2] R. B. Penman, T. Baldwin and D. Martinez, "Web Scraping Made Simple with SiteScraper," *Citeseer*, pp. 1-10.

[3] Wikipedia. *Web scraping*. (May. 18, 2022). Accessed: May. 18, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Web_scraping

[4] W. Roush. (2012, Jul 25). *Diffbot Is Using Computer Vision to Reinvent the Semantic Web*. [Online]. Available: https://xconomy.com/san-francisco/2012/07/25/diffbot-is-using-computer-vision-to-reinvent-the-semantic-web/

[5] Pinsent Masons (Out-Law News). *Google thumbnails are fair use, says Court of Appeals*. (May. 18, 2007). Accessed: May. 18, 2022. [Online]. Available: https://www.pinsentmasons.com/out-law/news/google-thumbnails-are-fair-use-says-court-of-appeals

[6] The Electronic Frontier Foundation (EFF). *Perfect 10 v. Google*. (May. 16, 2007). Accessed: May. 18, 2022. [Online]. Available: https://www.eff.org/cases/perfect-10-v-google

[7] J. Schultz. (2007, May 16). *P10 v. Google: Public Interest Prevails in Digital Copyright Showdown*. [Online]. Available: https://www.eff.org/deeplinks/2007/05/p10-v-google-public-interest-prevails-digital-copyright-showdown

[8] Pinsent Masons (Out-Law News). *Google image search results do not infringe copyright, says German court*. (Apr. 30, 2010). Accessed: May. 18, 2022. [Online]. Available:

https://www.pinsentmasons.com/out-law/news/google-image-search-results-do-not-infringe-copyright-says-german-court

[9] E. N. Sarr, S. A. L. L. Ousmane and A. Diallo, "FactExtract: automatic collection and aggregation of articles and journalistic factual claims from online newspaper". *Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2018, pp. 336-341. IEEE.

[10] S. Ashouri et al., "Indicators on firm level innovation activities from web scraped data," *Data in Brief*, 108246, 2022.

[11] A. Hajikhani et al., "Connecting firm's web scraped textual content to body of science: Utilizing microsoft academic graph hierarchical topic modeling," *MethodsX*, vol. 9, no. 101650, 2022.

[12] U. Baskaran and K. Ramanujam, "Automated scraping of structured data records from health discussion forums using semantic analysis," *Informatics in Medicine Unlocked*, vol. 10, pp. 149-158, 2018.

[13] R. A. Melchor et al., "CT-152: Application of Web-Scraping Techniques for Autonomous Massive Retrieval of Hematologic Patients' Information During SARS-CoV2 Pandemic," *Clinical Lymphoma Myeloma and Leukemia*, vol. 20, pp. 214, 2020.

[14] M. F. C. Portugal et al., "Epidemiological Analysis of 5,595 Procedures of Endovascular Correction of Isolated Descending Thoracic Aortic Disease Over 12 Years in the Public Health System in Brazil," *Clinics*, vol. 76, 2021.

[15] M. J. Lee, J. Kang, K. Hreha and M. Pappadis, "A Novel Web Scraping Approach to Identify Stroke Outcome Measures: A Feasibility Study," *Archives of Physical Medicine and Rehabilitation*, vol. 103(3), pp. 30, 2022.

[16] S. Mohan, A. K. Solanki, H. K. Taluja and A. Singh, "Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach," *Computers in Biology and Medicine*, vol. 144, no. 105354, 2022.

[17] L. Cui, Z. Jiang, X. Huang, S. Liu, Y. Wu and M. Fan, "Decade changes of the food web structure in tropical seagrass meadow: Implication of eutrophication effects," *Marine pollution bulletin*, vol. 173, no. 113122, 2021.

[18] Q. Wang, S. Fu, F. Mu, Z. Zhang and X. Liu, "Bottom aquaculture can improve the basic trophic pathways and enhance the secondary production: Implications from benthic food web analysis," *Marine Pollution Bulletin*, vol. 177, no. 113562, 2022.

[19] C. Muehlethaler and R. Albert, "Collecting data on textiles from the internet using web crawling and web scraping tools," *Forensic Science International*, vol. 322, no. 110753, 2021.

[20] M. Klasson, C. Zhang and H. Kjellström, "Using Variational Multi-view Learning for Classification of Grocery Items," *Patterns*, vol. 1(8), no. 100143, 2020.

[21] M. Kiran and N. Mownika, "Machine learning integrated emotions detection on lockdowns in India using advanced web scraping," *Materials Today: Proceedings*, 2021.

[22] J. Maybir and B. Chapman, "Web scraping of ecstasy user reports as a novel tool for detecting drug market trends," *Forensic Science International: Digital Investigation*, vol. 37, no. 301172, 2021.

[23] J. Schedlbauer, G. Raptis and B. Ludwig, "Medical informatics labor market analysis using

web crawling, web scraping, and text mining," *International Journal of Medical Informatics*, vol. 150, no. 104453, 2021.

[24] L. Ricci et al., "Web-based and machine learning approaches for identification of patient-reported outcomes in inflammatory bowel disease," *Digestive and Liver Disease*, vol. 54(4), pp. 483-489, 2022.

[25] T. Alipourfard, H. Arefi and S. Mahmoudi, "A Novel Deep Learning Framework by Combination of Subspace-Based Feature Extraction and Convolutional Neural Networks for Hyperspectral Images Classification," *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 4780-4783.

[26] T. Hayit, H. Erbay, F. Varçın, F. Hayit and N. Akci, "Determination of the severity level of yellow rust disease in wheat by using convolutional neural networks," *Journal of Plant Pathology*, vol. 103(3), pp. 923-934, 2021.

[27] W. Guo, G. Xu, B. Liu and Y. Wang, "Hyperspectral Image Classification Using CNN-Enhanced Multi-Level Haar Wavelet Features Fusion Network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022.

[28] P. Aggarwal, N. K. Mishra, B. Fatimah, P. Singh, A. Gupta and S. D. Joshi, "COVID-19 image classification using deep learning: Advances, challenges and opportunities," *Computers in Biology and Medicine*, no. 105350, 2022.

[29] T. Hayıt ve G. Çınarer, "X-RAY görüntülerini kullanarak GLCM ve derin özniteliklerin birleşimine dayalı Covid-19 sınıflandırılması," *İnönü Üniversitesi Sağlık Hizmetleri Meslek Yüksek Okulu Dergisi*, c. 10 (1), ss. 313-325, 2022

[30] K. Adem, S. Kiliçarslan and O. Cömert, "Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification," *Expert Systems with Applications*, vol. 115, pp. 557-564, 2019

[31] S. Kiliçarslan and M. Celik, KAF+ RSigELU: a nonlinear and kernel-based activation function for deep neural networks," *Neural Computing and Applications*, pp. 1-15, 2022

[32] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[33] M. Hussain, J. J. Bird and D. R. Faria, "A study on cnn transfer learning for image classification," in *UK Workshop on computational Intelligence*, Springer, Cham ,2018, pp. 191-202.

[34] Anonimous. *The Selenium Browser Automation Project*. (Mar. 16, 2022). Accessed: Apr. 12, 2022. [Online]. Available: https://www.selenium.dev/documentation/

[35] Anonimous. *ChromeDriver*. Accessed: Apr. 12, 2022. [Online]. Available: https://chromedriver.chromium.org/home

[36] S. P. Mohanty, D. P. Hughes and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in plant science*, vol. 7, no. 1419, 2016.

[37] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[38] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[39] C. Szegedy et al., “Going deeper with convolutions,” *IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[40] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, “Densely connected convolutional networks,” *IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.